Neural evidence for a distinction between short-term memory and the focus of attention

Lewis-Peacock, Jarrod A ; Drysdale, Andrew T ; Oberauer, Klaus ; Postle, Bradley R

Abstract: It is widely assumed that the short-term retention of information is accomplished via maintenance of an active neural trace. However, we demonstrate that memory can be preserved across a brief delay despite the apparent loss of sustained representations. Delay period activity may, in fact, reflect the focus of attention, rather than STM. We unconfounded attention and memory by causing external and internal shifts of attention away from items that were being actively retained. Multivariate pattern analysis of fMRI indicated that only items within the focus of attention elicited an active neural trace. Activity corresponding to representations of items outside the focus quickly dropped to baseline. Nevertheless, this information was remembered after a brief delay. Our data also show that refocusing attention toward a previously unattended memory item can reactivate its neural signature. The loss of sustained activity has long been thought to indicate a disruption of STM, but our results suggest that, even for small memory loads not exceeding the capacity limits of STM, the active maintenance of a stimulus representation may not be necessary for its short-term retention.

DOI: https://doi.org/10.1162/jocn_a_00140

Published in final edited form as:
J Cogn Neurosci. 2012 January ; 24(1): 61–79. doi:10.1162/jocn_a_00140.

Neural Evidence for a Distinction Between Short-Term Memory and the Focus of Attention

Jarrod A. Lewis-Peacock, **Andrew T. Drysdale**, **Klaus Oberauer**, and **Bradley R. Postle**

Abstract

It is widely assumed that the short-term retention of information is accomplished via maintenance of an active neural trace. However, we demonstrate that memory can be preserved across a brief delay despite the apparent loss of sustained representations. Delay-period activity may in fact reflect the focus of attention, rather than short-term memory. We unconfounded attention and memory by causing external and internal shifts of attention away from items that were being actively retained. Multivariate pattern analysis of fMRI indicated that only items within the focus of attention elicited an active neural trace. Activity corresponding to representations of items outside the focus quickly dropped to baseline. Nevertheless, this information was remembered after a brief delay. Our data also show that refocusing attention towards a previously unattended memory item can reactivate its neural signature. The loss of sustained activity has long been thought to indicate a disruption of short-term memory, but our results suggest that, even for small memory loads not exceeding the capacity limits of short-term memory, the active maintenance of a stimulus representation may not be necessary for its short-term retention.

INTRODUCTION

Since at least the time of Hebb (1949) it has widely been assumed that the short-term retention of information is accomplished via maintenance of an active memory trace. This view has been reinforced by reports of elevated delay-period activity in extracellular (Fuster & Alexander, 1971; Kubota & Niki, 1971), electroencephalographic (Vogel, McCollough, & Machizawa, 2005), and hemodynamic (Courtney, Ungerleider, Keil, & Haxby, 1997; Haxby, Petit, Ungerleider, & Courtney, 2000; Curtis & D'Esposito, 2003) recordings of animals and humans. Consequently, the loss of sustained activity is thought to indicate a disruption of the memory trace (di Pellegrino & Wise, 1993; Miller & Desimone, 1994; Postle, Druzgal, & D'Esposito, 2003). However, to the best of our knowledge, virtually all studies of the short-term retention of information (regardless of species, procedure, concurrent physiological measurement, etc.) have confounded *memory* with *attention*: The information to be remembered is the most task-relevant information throughout the memory interval, and therefore is likely to be continuously attended to. This leaves open the question of whether sustained delay-period activity is better understood as a correlate of memory, or as a correlate of attention. To address this question, we unconfounded these constructs across two experiments by causing external and internal shifts of attention away from information that was being actively retained during a brief memory interval. Using multivariate pattern analysis (MVPA) of brain activity recorded in event-related fMRI (Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009), we tested the hypothesis that delay-period activity reflects the information that is being attended to, but not the information that is unattended, yet remembered, after a brief delay. The *embedded-component* theory of information processing provides the theoretical framework for this hypothesis. It characterizes short-term memory (STM) as an

emergent property of the interaction of long-term memory (LTM) and attention (Cowan, 1988; Cowan, 1995; Ericsson & Kintsch, 1995; Oberauer, 2002), and postulates a distinction between a capacity-limited central component of STM, referred to as the *focus of attention*[1], and a more peripheral component referred to as *activated LTM*. In keeping with this view, we use the term STM to refer not to a hypothetical system, but to the *ability* of the mind/brain to retain a limited amount of information over brief periods of time.

This model accounts for a wide range of data from behavioral, neuropsychological, electrophysiological, and neuroimaging studies of monkeys and humans (reviewed in Postle, 2006). For example, evidence for the interaction between attention and LTM comes from electroencephalographic recordings of increased neural synchrony between prefrontal and posterior cortices during STM (Ruchkin, Grafman, Cameron, & Berndt, 2003). This observation has motivated the idea that prefrontal cortex directs the attentional focus needed for maintaining activation in the appropriate posterior processing regions. Initial neuroscientific support for engagement of LTM in STM relied on demonstrations that the brain regions which participate in the initial perception and comprehension of incoming information are also involved in its short-term retention. For example, delay-period activity during STM for faces has been localized to regions of temporooccipital cortex that are believed to support the perception and long-term retention of faces (Druzgal & D'Esposito, 2003; Postle et al., 2003; Ranganath, Cohen, Dam, & D'Esposito, 2004; Ranganath, DeGutis, & D'Esposito, 2004). Such results cannot be interpreted as strong tests of this model, however, because they rely on tenuous reverse inferences (i.e., they reason backwards from the presence of peaks in brain activity to the engagement of a particular cognitive function, Poldrack, 2006). This is because, for example, the presence of sustained activity peaks in mid-fusiform gyrus does not necessarily imply that faces were being remembered, because this region can show above-baseline activity during many other cognitive states (e.g., Gauthier, Skudlarski, Gore, & Anderson, 2000). Stronger evidence comes from a demonstration with MVPA that the information content of delay-period activity can be decoded based on distributed patterns of unthresholded brain activity recorded from an independent LTM task (Lewis-Peacock & Postle, 2008). MVPA can support stronger reverse inferences than univariate techniques because it captures high-dimensional neural representations that have markedly higher selectivity than do univariate activation peaks, a consequence of which is that MVPA can support discrimination of neural representations at the item level (Kriegeskorte, Formisano, Sorger, & Goebel, 2007).

The temporal dynamics of the embedded-component model are being mapped out in the behavioral literature. For example, memory items that are no longer relevant for behavior can be removed (within 1–2 s) from the focus of attention, thereby reducing the load on the system's limited capacity and consequently reducing response times to memory probes of the behaviorally-relevant items still in the focus (Oberauer, 2001). Information removed from the focus remains in a state of heightened availability for several seconds, as shown by the finding that lures from a recently encoded memory list are harder to reject than lures not recently encountered (Monsell, 1978; Woltz, 1996; Oberauer, 2001). This information can be re-focused if needed again (Oberauer, 2005), otherwise it is prone to forgetting by decay or by interference.

A recent fMRI study showed that retention of a single item inside the focus of attention exhibits a distinct neural signature (Nee & Jonides, 2008). It found that an item within the

[1]What we refer to as the "focus of attention" is the broad focus of attention (Cowan, 1995) which has a capacity limit of about four items. This contrasts with a narrow focus of attention, consisting of a single item, that is differentiated from the "direct access region" while can hold about four items (Oberauer, 2002). Our data do not address the distinction between these constructs, and therefore we consistently imply the broader definition.

focus is associated with increased activation in inferior temporal cortex relative to other information in STM. Attended information was sustained via enhanced functional connectivity with frontal and posterior parietal regions, whereas unattended information was characterized by increased activations in LTM retrieval-related regions in the medial temporal lobe and prefrontal cortex. These intriguing results provide some of the first empirical evidence for a neural dissociation of representations within STM. Two aspects of the present study gave it the potential to provide novel insight into the embedded-component model. First, it used MVPA so that, rather than having to make assumptions about what elevated activity in one or more brain regions might represent, we could objectively, quantitatively measure the information being actively represented during the delay period. Second, we explicitly unconfounded attention from STM by exogenously and endogenously drawing the focus of attention away from information that had to be remembered after a brief delay. In the first experiment, we recorded fMRI data from healthy young adults while they performed a paired-associate recognition test of STM in which, during an unpredictable half of trials, trial-irrelevant stimuli were presented in the middle of a memory delay. These visual distractors were used to redirect the focus of attention outwardly towards external stimuli and away from the items being actively retained in memory. In the second experiment, we recorded fMRI data from a separate group of participants while they performed a test of STM during which only one of two items being actively retained in STM was cued as relevant for the next behavioral response. These cues were used to redirect (i.e., shrink) the focus of attention internally such that the irrelevant item would be removed from the focus.

Our results showed that the information content of delay-period activity reflects the focus of attention rather than the full contents of STM. In fact, brain activity corresponding to representations of unattended information dropped to baseline levels Nevertheless, this information was remembered after a brief delay. Our data also showed that re-focusing attention to previously unattended information can restore the active neural signature of that information. Whereas the loss of sustained activity has been thought to indicate a disruption of STM, our results suggest that active maintenance may not be required for the short-term retention of information. Instead, two complementary forms of retention may underlie STM: (1) the active retention of information *inside* the focus of attention via sustained neural firing, and (2) the passive retention of information *outside* the focus via some other neural mechanism (e.g., transient changes in synaptic potentiation) from which it can be reactivated with cue-based retrieval. The present results provide direct demonstrations of the former, and they demand the latter by inference. Theoretically, our results call for rethinking the "activation" assumption for memory representations outside the focus of attention in the embedded-component model. Empirically, they suggest that many previous studies of short-term and working memory might best be interpreted as studies of sustained attention to information.

EXPERIMENT 1

Methods

Participants—Fourteen (9 men; ages 18–29) healthy right-handed adults were recruited from the undergraduate and medical campuses of the University of Wisconsin-Madison. None reported any medical, neurological, or psychiatric illness, and all gave informed consent. One participant's data were removed from analysis due to a failure to comply with task instructions.

Phase 1: Short-Term Recognition—Participants performed short-term recognition of a total of 120 pictures selected from three categories – 40 unfamiliar faces (20 males, 20 females), 40 unfamiliar outdoor places/scenes, and 40 common objects (Fig. 1A). All

images were converted to greyscale with image processing software to remove any unintended confounds of color in the perception and short-term retention of the stimuli. Each stimulus was presented one time only, for a total of 120 randomly ordered stimulus presentations. Each trial consisted of a target presentation (1 s), a delay period (7 s), and a probe presentation (1 s). Participants indicated with a Yes/No button press whether the probe stimulus matched the target stimulus. Trials were configured such that there was a probability of 0.5 that the probe stimulus was the same as the target, with foils (invalid probes) drawn from the same category as the target. The inter-trial interval period (ITI) lasted 13 s and consisted of an arithmetic task (7 s), requiring evaluation of the sum of three numbers, a task intended to reduce interference between trials and encourage alertness throughout the experiment (Polyn, Natu, Cohen, & Norman, 2005; Lewis-Peacock & Postle, 2008), and a final rest period (6 s) before the next trial began.

Phase 2: Stimulus Pairing—Ranging from 0 to 42 days following their initial scan, participants returned to complete Phases 2 and 3 of the experiment. For Phase 2, which occurred outside the scanner, 18 stimuli (six faces, six places, and six objects) were selected at random (a different subset for each participant) from the initial set and paired arbitrarily so that nine stimulus pairs were created (Fig. 1B). Each pair consisted of two stimuli from different categories *face-place*, *face-object*, and *place-object* pairs). Participants learned these pairings via repeated three-alternative forced choice testing (with foils drawn from the set of 18) until they achieved a criterion-level performance of 72 consecutive correct trials. The learning task was completed in approximately five minutes for each participant.

Phase 3: Short-Term Paired-Associate Recognition—Immediately after learning the stimulus pairs, participants returned to the scanner and performed paired-associate recognition with those stimuli (Fig. 1C). Each trial consisted of a target stimulus (1 s), a delay period (11 s), a probe stimulus (1 s), and a rest period between trials (13 s). Participants indicated with a Yes/No button press whether the probe stimulus was the correct associate of the target stimulus. Trials were configured such that there was a probability of 0.5 that the probe stimulus was the correct associate of the target, with foils drawn from the trial-irrelevant category (i.e., the category to which neither the target nor its associate belonged). The trial depicted in Fig. 1C is an example of a *face-place* trial: the target was a face and its paired-associate stimulus was a place. Randomly, on half of the trials, four trial-irrelevant "distractor" pictures were presented during the delay period in rapid succession (0.5 s per stimulus, 2 s total). These stimuli were always selected from the trial-irrelevant category (e.g., object stimuli on a *face-place* trial). Participants passively observed these stimuli and were instructed not to divert their attention from the center of the screen when they appeared. There were a total of 144 trials (72 with distraction, 72 without distraction). One third (i.e., 48) of the trials involved *face-place* pairs, one third involved *face-object* pairs, and the remaining one third involved *object-place* pairs. For each pair, half of the trials presented one stimulus as the target (e.g., the face stimulus from a *face-place* pair), and the other half of the trials presented its associate as the target (e.g., the place stimulus from the same *face-place* pair). Each of the 18 unique pairs was presented in a total of eight trials (four times in each direction). (Note that although this task requires LTM for stimulus pairings, it is a test of STM, because the correct evaluation of the probe requires memory for what was presented at the beginning of the trial.)

Cognitive Strategies—In our previous study (Lewis-Peacock & Postle, 2008), we observed large variability in the cognitive strategy employed by our participants to solve a short-term paired-associate recognition task. Some participants favored a *retrospective* strategy (i.e., they thought about the stimulus that was presented at the beginning of the trial), others favored a *prospective* strategy (i.e., they retrieved from LTM the associate of

the stimulus that was presented and thought about it for the remainder of the delay period), and still others switched between the two strategies across trials. In the Phase 3 task of Experiment 1 in the present study, we attempted to control for variability in strategies by instructing half of our participants to use a retrospective strategy on every trial ("hold the first picture in mind and try not to think about its associate until the probe appears"), and the other half to use a prospective strategy ("as soon as you see the first picture, quickly recall its associate and hold it in mind"). This manipulation was designed to allow the independent observation of the effects of distraction on representations derived from visual perception (in participants using the retrospective strategy) and on representations recalled from LTM (in participants using the prospective strategy). In accordance with findings in the monkey (Takeda, Naya, Fujimichi, Takeuchi, & Miyashita, 2005), we predicted that the neural representation in inferotemporal cortex of the target stimulus, but not its associate, would be disrupted by the distractors. Assuming that active neural representation is the neural basis for STM, one would predict that the loss of the target representation would cause the participant to forget, and thus be forced to guess about the validity of the memory probe, with a consequent decline in behavioral performance.

Data Collection—All tasks were implemented with E-Prime software version 2.0 (Psychology Software Tools), and an Avotec goggle system (Avotec, Inc., Stuart, Florida) was used to display visual stimuli inside the scanner. Whole-brain images were acquired with a 3-T scanner (GE Signa VH/I). For all participants, we acquired high-resolution T1-weighted images (30 axial slices, 0.9375 × 0.9375 × 4 mm). We used a gradient-echo, echo-planar sequence (time repetition = 2000 ms, echo time = 50 ms) to acquire data sensitive to the blood oxygen level-dependent (BOLD) signal within a 64 × 64 matrix (30 axial slices coplanar with the T1 acquisition, 3.75 × 3.75 × 5 mm). Eight blocks of the Phase 1 short-term recognition task were obtained, each scan consisting of 15 trials lasting 5 min 50 s, for a total of 46 min 40 s in functional scans. All task runs were preceded by 20 s of dummy pulses to achieve a steady state of tissue magnetization. Eight blocks of the Phase 3 paired-associate recognition task were also obtained, each scan consisting of 18 trials lasting 8 min 8 s, for a total of 65 min 4 s in functional scans. Across both tasks, each participant was tested for a total of 111 min 45 s.

Preprocessing—Preprocessing of the functional data was done with the AFNI software package using the following preprocessing steps, in order: correction for slice time acquisition and rigid-body realignment to the first volume from the experimental task with 3dvolreg; removal of signal spikes with 3dDespike; removal of the mean from each voxel and linear and quadratic trends from within each run with 3dDetrend; and correction for magnetic field inhomogeneities (using in-house software). Finally, functional data from the second task were aligned to data from the first task using 3dAllineate. Note that neither was spatial smoothing imposed, nor were the data spatially transformed into a common atlas space prior to hypothesis testing. Rather, the data from each participant were analyzed in that participant's un-smoothed, native space.

For classification analyses, a feature selection analysis of variance (ANOVA) was applied to the preprocessed images to select those voxels whose activity varied significantly ($p<0.05$) between face, place, and object categories over the course of the Phase 1 task. This standard procedure reduces noise in the classification analyses by removing uninformative voxels. (Note: We repeated the analyses reported here without prior feature selection, which produced qualitatively similar, though quantitatively noisier, results.) The number of voxels passing feature selection was 4,540 (s.d. 2,255). Searchlight classification analyses (with a sphere radius of 2 (7 total voxels), 3 (19 total voxels), or 4 (33 total voxels); see Kriegeskorte, Goebel, & Bandettini, 2006) were also applied to the Phase 1 data to assess the extent of category-specific information throughout the brain. Classifier decoding of

Phase 3 data using voxels selected by the searchlight technique produced qualitatively similar results to those selected by the simpler ANOVA procedure, and therefore only results from the ANOVA-based feature selection masks are reported. Many previous accounts have emphasized the importance of prefrontal cortex (PFC) in supporting the temporary retention of information across distraction. To address this idea, we divided the feature-selected voxels into "no-PFC" and "PFC-only" masks. Anatomically-derived PFC masks were generated for each participant in AFNI by backwards transforming a TT_Daemon atlas mask (consisting of Brodmann areas 8–11, 44–46) into that participant's native space. New "no-PFC" masks were created by removing all PFC voxels from the original feature-selected set. The number of voxels retained in each condition was 3,844 (s.d. 1,908) for the "no-PFC" condition, and 696 (s.d. 347) for the "PFC-only" condition. An additional mask was created for each participant covering the inferotemporal cortex (ITC), which consisted of the inferior temporal, middle temporal, and fusiform gyri (403 voxels, s.d. 156). These masks were created in a similar fashion as the PFC masks. Voxels from these masks served as input nodes to the pattern classifier for hypothesis testing.

Multivariate Pattern Analysis: Training—A pattern classifier was trained, separately for each participant, on data from the delay-period of the Phase 1 task. The Princeton Multivoxel Pattern Analysis Toolbox (MVPA, http://code.google.com/p/princeton-mvpa-toolbox), in conjunction with the Matlab Neural Network Toolbox, was used for all classification analyses (see Haynes & Rees, 2006; Norman et al., 2006; Pereira et al., 2009 for reviews). Data from the initial 8 s, at intervals of TR = 2 s, of each trial from Phase 1 were used to train a two-layer (no hidden layers) feedforward neural network via Matlab's *trainscg* scaled conjugate gradient backpropagation algorithm, with sigmoidal transfer functions between the input layer (N voxels) and output layer (3 stimulus categories) of the network. The classifier was trained to distinguish patterns of brain activity corresponding to the short-term retention of faces, places, and objects. Note that data from the ITI was not used as a baseline in training because the interval between trials was filled with a secondary task (arithmetic) that engaged the brain more strongly than is characteristic of an unfilled ITI (see Experiment 2). To assess empirically the inclusion of the first TR of each trial (during which the visual stimulus was on screen for the first 1 s) we calculated the classification accuracy at each time interval of the 8 s training window and found that category discrimination was well above chance throughout the entire period. Thus, we are confident that comparable stimulus category-specific activity was being evoked throughout the first 8 s of the trial, despite contamination from the initial perception and encoding of the target stimulus. A unique classifier was created for each participant and applied only to that participant's data. To reduce prediction error in analyses involving the non-deterministic backpropagation classifier algorithm, the reported results were the average of 50 network iterations, each initialized with a different set of random weights. All data used to train the classifiers were shifted back in time by 4 s to account for hemodynamic lag of the BOLD signal. Therefore, the 8 s of fMRI data that were used from each trial were actually data that were recorded between 4 and 12 s after the beginning of the trial. This adjustment, although crude, reasonably accommodates the slow hemodynamic response and is standard practice in multivariate pattern analysis. As a check on validity, we retrained the classifier using a 6 s lag adjustment, and this did not significantly alter the results. We evaluated classifier training accuracy by using the method of k-fold cross-validation, i.e., training on k-1 blocks of data and testing on the k^{th} block, and then rotating and repeating until all trials had been classified. For each 2-sec TR of fMRI data, the classifier produced an estimate (from 0 to 1) of the extent to which the brain activity matched the pattern of activity corresponding to the three categories it had been trained on. These estimates reflected the classifier's *evidence* for each category. The classifier's *prediction* at each TR corresponded to the category with the most evidence.

Prediction accuracy was calculated as the proportion of TRs in which the classifier correctly predicted the actual category of the trial from which that TR was sampled.

To assess the relative importance of different brain areas to the classification of the stimulus categories, we determined, from the trained pattern classifier, which voxels were important for identifying patterns of brain activity corresponding to each of the three categories. We applied the voxel importance formula (from Polyn et al., 2005): $imp_{ij} = 100 * w_{ij} * avg_{ij}$, where w_{ij} is the weight between input unit *i* and output unit *j*, and avg_{ij} is the average activity of input *i* during the short-term retention of category *j*. Importance maps for the three categories were calculated for each participant, transformed into standardized space using AFNI's *@auto_tlrc* and *adwarp*, blurred with a full-width half-max of 4 mm and averaged across all participants with *3dmerge*, thresholded at an importance score of 2.0, and overlaid on an inflated anatomical version of the N27 brain dataset (Holmes et al., 1998) using AFNI's surfacing mapping utility (SUMA) for display purposes.

Multivariate Pattern Analysis: Testing—A trained pattern classifier for each participant, trained on all eight blocks of Phase 1 data, was used to assess the extent to which category-specific patterns of brain activity reappeared during the delay-period of the Phase 3 task. Preprocessed fMRI data at intervals of TR = 2 s were classified from the initial 20 s of each trial (Fig. 1C), corresponding to target presentation (1 s), delay period (11 s), probe presentation (1 s), and the first 7 s of the ITI (which was not rest, but filled with an arithmetic task). Pattern classification of these data allowed us to distinguish brain activity corresponding to the target, its associate, and the trial-irrelevant category. If, for example, a *face*-like delay-period activity pattern was identified on a *face-place* trial, this would suggest that the brain was actively maintaining, via persistent brain activity, a representation of the face stimulus presented at the beginning of the trial, consistent with a retrospective strategy. Delay-period activity reflecting a prospective strategy would consist of brain activity patterns identified as corresponding to the category of the target's associate (in the example, *places*). This could only occur if, upon seeing the target stimulus, the participant retrieved from LTM the representation of its associate and actively retained this representation. The amount of distraction-induced brain activity during the delay period would be indicated by the classifier's evidence for the category of the distractors (in the example, *objects*). Importantly, the continuous decoding of data from these trials allowed for a complete characterization of the evolution of category-specific representations throughout each trial, allowing for the detection of transitions between target-, distractor-, and associate-related activity within the same brain regions. Note that possible contamination of delay-period activity due to perceptual processing of the probe stimulus was not a concern, as this processing would be expected to introduce noise, not coherent category-specific activity. This follows from the fact that the stimulus presented as the probe was from the same category as the associate of the target on only half of the trials, the remaining trials presented foils drawn from a different category.

Searching for Distraction Resistance—An additional analysis was designed to search the brain for any evidence of distraction-resistant STM representations. The purpose of this analysis was to identify voxels whose activity in the Phase 3 task, after being decoded by the classifier, would show that a task-relevant stimulus representation was sustained in the face of distraction. We selected voxels whose activity appeared to be the least responsive to presentation of the distractors, and then assessed whether decoding the brain activity from these regions produced interpretable and reliable evidence of distraction-resistance. If this analysis failed, we reasoned that it would be unlikely to find such representations anywhere else in the brain. We applied a modified version of the searchlight classification technique (Kriegeskorte et al., 2006). To search for distraction-resistant activity in the prospective strategy group, we identified spheres of voxels (separately using a radius of 2, 3, or 4

voxels) that both (1) coded for the associate stimulus and (2) were least responsive to the distractors. We recorded, for all spheres, the proportion of post-distraction data (i.e., data from distraction-present trials between the onset of distraction and the onset of the probe, 6–12 s) during which the classifier's evidence for the associate's category was higher than its evidence for all other categories. This proportion was assigned to the center voxel of the sphere, and then the sphere was shifted and this procedure was repeated until all spheres had been tested. A complementary algorithm implemented a search for distraction-resistant activity for the target stimulus in the retrospective strategy group. The resulting statistical voxel maps were thresholded (at scores of 0.45) using estimates from a χ^2 distribution test with $df = 2$, using a strict alpha of 2×10^{-6} as a Bonferroni correction for multiple comparisons. (Note: These maps were also thresholded using an uncorrected alpha, which produced qualitatively similar results.) Voxels from all supra-threshold spheres were combined into one mask and used as input to the classifier for retraining on Phase 1 and retesting on Phase 3. For spheres of radius 3, the average number of voxels in prospective strategy group was 240 (s.d. 243), and the average number of voxels in the retrospective strategy group was 190 (s.d. 56).

Results (Phase 1)

Behavior—The mean accuracy and response time across all participants in the Phase 1 task were 94% (SEM 1) and 650 ms (SEM 10). Response times from trials with an incorrect response were excluded. A three-way repeated measures ANOVA on response accuracy with stimulus category (*face/place/object*) as a within-subjects factor revealed a significant main effect of stimulus category ($F(2,24)=3.50$, $p=0.046$), and follow-up pairwise comparisons (with Bonferroni correction) indicated that the accuracy on object trials (96 %, SEM 1) was marginally higher ($p=0.053$) than the accuracy on place trials (91%, SEM 2). An identical ANOVA on response time revealed a significant main effect of stimulus category ($F(2,24)=9.36$, $p<0.001$), but follow-up pairwise comparisons (both with or without Bonferroni correction) indicated that there were no reliable differences between any category pairs.

MVPA—Brain data from all Phase 1 trials were used to train a classifier separately for each participant. Group-averaged classification performance showed that brain activity from the delay period of the Phase 1 task was reliably classified as consistent with the appropriate category of the trial (Fig. 2A). The classifier's prediction accuracy for each category was significantly above chance (33 %) based on one-tailed, independent-sample t-tests across participants, with $p<0.005$, for all three categories. The mean classifier evidence for each category showed strong category selectivity (e.g., the *face* evidence was selectively high for face trials), supported by a significant interaction of trial type (*face/place/object*) × evidence type (*face/place/object*) from a 3×3 repeated measures ANOVA on the classifier evidence values ($F(4,48)=220.09$, $p<0.001$). For clarity, only data from the "no-PFC" condition are shown here. However, training the classifier on voxel activity from the whole brain, or from voxels restricted only to PFC or ITC, was also successful (but performance in PFC was considerably closer to chance-level prediction than in the other regions). Although established category-selective areas contributed to the classification of the three categories (e.g., the mid-fusiform gyrus for faces, the parahippocampal gyrus for places, and the lateral occipital cortex for objects), multiple, distributed brain regions were also identified as important for each category (Fig. 2B). This replicates previous findings when famous faces, famous places, and common objects were evaluated in a test of LTM (Polyn et al., 2005; Lewis-Peacock & Postle, 2008).

Results (Phase 3)

Behavior—The mean accuracy and response time across all participants in the Phase 3 task were 96% (SEM <1) and 778 ms (SEM 11). A 2×2×6 mixed ANOVA on response accuracy, with instructed strategy (*retrospective/prospective*) as a between-subjects factor and distraction condition (*absent/present*) and trial type (6 pairwise combinations of *face, object, & scene*) as within-subjects factors, revealed a marginally significant main effect of instructed strategy ($F(1,11)=4.18$, $p=0.065$), indicating a trend that the prospective strategy (98%, SEM <1) produced better accuracy than the retrospective strategy (95%, SEM 1). The main effect of distraction was also marginally significant ($F(1,11)=4.36$, $p=0.061$), indicating a trend that participants responded more accurately to distraction-present trials (97%, SEM 1) than to distraction-absent trials (96%, SEM 1). However, neither of these main effects were statistically reliable at the standard alpha=5% cutoff. The main effect of trial type ($F(1,11)=0.35$, $p=0.882$) and all interactions between the factors were non-significant. An identical 2×2×6 mixed ANOVA on response times revealed a significant main effect of distraction condition ($F(1,11)=46.86$, $p<0.001$), indicating that participants responded faster on trials with distraction (734 ms, SEM 14) than on trials without distraction (823 ms, SEM 16). This difference likely reflected a general attentional enhancement for distraction-present trials due to the processing of additional stimuli during the otherwise long, unfilled delay period (see also Postle, Idzikowski, Della Salla, Logie, & Baddeley, 2006). A related possibility is that because the distractors were always from a different category than the target and its associate, the presentation of distractors during the delay period may have served to reduce uncertainty about the category of the target's associate, thus narrowing the retrieval space and facilitating performance. The main effect of trial type was significant ($F(5,55)=2.44$, $p=0.045$), but follow-up pairwise comparisons (both with and without Bonferroni correction) revealed no reliable differences between any pair of trial types. The main effect of instructed strategy ($F(1,11)=1.92$, $p=0.193$), and all interactions between factors were non-significant.

MVPA—Brain data from all Phase 3 trials were decoded, separately for each participant, using a classifier that was trained on data from all Phase 1 trials. For clarity, we present only results from the "no-PFC" region of interest[2]. For participants who were instructed to retain the perceptual stimulus during the delay ("retrospective strategy"), sustained representation of this stimulus were identified on distraction-absent trials, as indicated by relatively greater evidence for the target category throughout the delay (Fig. 3, top left). Although strong evidence for the target category was also observed during the early portion of the delay period on distraction-present trials, it was sharply attenuated and replaced by evidence for the trial-irrelevant category following the onset of the distractors (Fig. 3, bottom left). This result indicates that the active neural representation of the target stimulus (as assessed by MVPA) was replaced by perceptual representations of the distractors. For participants instructed to retrieve the target's associate and retain it in anticipation of the probe ("prospective strategy"), sustained representation of the category of the associate were identified on distraction-absent trials, indicated by a transition from strong evidence for the target to strong evidence for its associate during the delay (Fig. 3, top right). Because the probe stimulus had not yet been presented, any brain activity classified as consistent with the associate's category must have been reinstated from LTM. It has been proposed that information that is retrieved from LTM and then actively retained in STM is more robust to

[2]Decoding with voxels from the whole brain, or only those restricted to ITC, produced qualitatively similar results. However, although classifier training on Phase 1 data in PFC was successful, decoding of Phase 3 data from this region failed to produce interpretable results. PFC is thought to be a critical neural substrate for cognitive control and the representation of task demands (Miller & Cohen, 2001). Although the stimulus materials were identical between the training task (Phase 1) and testing task (Phase 3), the cognitive demands of each task were not (short-term recognition vs. short-term paired-associate recognition). This may underlie the classifier's inability to generalize from the training data to the testing data in PFC.

distraction than perceptually-derived information (Takeda et al., 2005). Contrary to this proposal, however, our results show that sustained category-specific information related to the LTM-derived associate stimulus was disrupted by the distractors. The classifier's evidence for the associate was attenuated (and became indistinguishable from the estimates of the task-irrelevant target stimulus) when distractors were presented during the delay, accompanied by a significant increase in evidence for the distractors (Fig. 3, bottom right).

A 2×2×3×10 mixed ANOVA on classifier evidence values with instructed strategy (*retrospective/prospective*) as a between-subjects factor and distraction condition (*absent/present*), stimulus type (*target/associate/irrelevant*), and time (*TRs 1–10*) as within-subjects factors revealed a significant three-way strategy × stimulus × time interaction ($F(18,198)=1.77$, $p=0.031$). This result supports the qualitative interpretation, suggested in Fig. 3, that task instruction had a differential effect on the trial-averaged classifier evidence values for the two groups of participants. The three-way distraction × stimulus × time interaction was also significant ($F(18,198)=11.11$, $p<0.001$), confirming that the distraction manipulation had a statistically reliable effect on the classifier evidence values across the duration of the trials. The four-way interaction of strategy × distraction × stimulus × time was non-significant ($F(18,198)=1.33$, $p=0.174$). Taken together, the results from both groups indicate that the active task-relevant representation was disrupted following distraction.

An additional analysis using a voxel searchlight technique identified, in each participant, a small set of voxels that exhibited a relatively weaker response to the distractor stimuli (see Methods). However, retraining a classifier on Phase 1 data from only these voxels failed to find any reliable evidence for distraction-resistant representations in the Phase 3 data (data not shown). Any brain region we tested that showed evidence of sustained representation of the task-relevant stimulus during the first half of the delay period also showed a robust neural response to the trial-irrelevant distractors, which in turn suppressed the activity pattern associated with the former. Therefore, despite applying two different classification approaches (from large regions of interest that included thousands of voxels, and from small searchlight spheres that included tens of voxels), we were unable to find any reliable evidence for distraction-resistant representations of trial-relevant information in the fMRI data.

Discussion

The effects of visual distraction during the delay period of the Phase 3 task were twofold: The pattern of distributed brain activity corresponding to a representation of the trial-relevant stimulus dropped to baseline, and yet there was no loss of recognition accuracy compared to trials without the distraction. This result is intriguing because classifier estimates of category-specific activity have been shown to accurately reflect the strength of neural representation of a specific stimulus (Newman & Norman, 2010; Quamme, Weiss, & Norman, 2010; Kuhl, Rissman, Chun, & Wagner, 2011). A strong interpretation of our results is that the short-term retention of information does not depend on persistent activation of representations of the remembered material. Two methodological issues that may cause concern with this interpretation are: (1) it is unclear whether the pattern classifier was capable of identifying multiple, concurrently active STM representations (if they existed), or whether the results merely reflected a winner-take-all classification outcome; (2) because the classifier was trained on delay-period activity from the Phase 1 data, it may have been unfair to directly compare decoding results for on-screen stimuli (the distractors) with decoding results for remembered stimuli (the targets and their associates), because perceptual stimulation engages the brain more strongly than does STM retention (Sheth & Shimojo, 2003; Serences, Ester, Vogel, & Awh, 2009). Experiment 2, however, was not susceptible to either of these concerns.

EXPERIMENT 2

Methods

Participants—Nine (5 men; ages 21–30) healthy right-handed adults were recruited from the undergraduate and medical campuses of the University of Wisconsin-Madison. None reported any medical, neurological, or psychiatric illness, and all gave informed consent.

Phase 1: Short-Term Recognition—Participants performed 72 trials of short-term recognition of a stimulus selected randomly from one of three categories -- English words, pronounceable pseudowords, and line segments -- with 24 trials drawn from each category (Fig. 4A). Each trial consisted of a category cue (2 s), a target presentation (0.5 s), a delay period (7.5 s), a probe presentation (0.5 s), a response period (1.5 s), followed by a blank screen (10 s) that preceded the next trial. Participants indicated with a button press whether the probe stimulus matched the item in memory according to a category-specific criterion. Trials were configured such that there was a probability of 0.5 that the probe stimulus satisfied the criterion. A synonym judgment was required for words, a rhyme judgment was required for pseudowords, and a visual orientation judgment was required for line segments. Foils (to-be-rejected probes) for the three categories were conceptually-unrelated words, single-syllable pseudowords with a non-matching vowel sound, and line segments in which one of the segments differed in orientation by at least 30 degrees. Although phonological, semantic, and visual encoding processes were likely involved in the processing of all memory items (Wickens, 1970), the stimuli and task were designed to encourage encoding in one primary domain of representation. That is, we attempted to elicit the short-term retention of information in a semantic (i.e., conceptual) form on trials that required a synonym judgment, in a phonological form on trials that required a rhyme judgment, and in a visual form on trials that required a line orientation judgment. Words were presented in white (on black background) to indicate that the stimulus was to be primarily encoded based on its semantic characteristics. Pseudowords were presented in cyan to indicate that the stimulus was to be primarily encoded based on its phonological characteristics. Line segments were always presented in white (on black background) and were to be primarily encoded in a visual form. The domain-specific comparison criteria used here were modeled after a rich literature highlighting dissociations between verbal and visual processes in STM (Baddeley, 1986), as well as more recent studies dissociating semantic and phonological components (Haarmann & Usher, 2001; Martin, Wu, Freedman, Jackson, & Lesch, 2003; Shivde & Thompson-Schill, 2004; Cameron, Haarmann, Grafman, & Ruchkin, 2005).

Phase 2: Short-Term Recognition with Relevance Cues—Participants performed a second short-term recognition task in the scanner immediately after completing the Phase 1 task. This task was modeled on a modified version of the Sternberg recognition task (Oberauer, 2005). At the beginning of each trial, one stimulus was presented on the top half of the screen, and another was presented on the bottom half (Fig. 4B). The two stimuli for each trial were always selected from separate categories such that two of the three stimulus categories were represented in every trial. Stimulus offset was followed by a brief delay, and then a cue indicating which memory item was relevant for the first recognition probe. The cues consisted of two inward-facing red arrows, centered on either the top or bottom half of the screen, the location of which corresponded to the location where a stimulus had been presented at the beginning of the trial. After the probe (and response), a second cue appeared which indicated the relevant memory item for a second recognition probe, with equal probability of cuing either item. Thus, until the onset of the second cue, both stimuli from the beginning of the trial needed to be retained for successful task performance. Trials in which the same memory item was selected by both cues are referred to as *repeat* trials, and the other trials are referred to as *switch* trials. Similar to the Phase 1 task, trials in Phase 2

were configured such that there was a probability of 0.5 that the probe stimulus satisfied the category-specific criterion, with foils chosen as before. There were 72 trials, one third of which involved stimuli representing each of the three combination of categories (i.e., words & pseudowords, words & lines, and pseudowords & lines).

Stimuli—Words were nouns, verbs, and adjectives selected from an online psycholinguistic database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm) with concreteness, imageability, and frequency of each within one standard deviation of the mean of the entire database. Pseudowords consisted of pronounceable single-syllable letter strings that were created for this study. Intended pronunciation of the pseudowords was based on standard English (i.e., a string ending with the letter ‘e’ indicated a long vowel sound and a string ending with a double consonant indicated a short vowel sound). No compound vowels (e.g., ‘ou’) were used. Line stimuli consisted of a pair of line segments, each line tilted between 10 to 170 degrees, at intervals of 10 degrees, away from vertical. Tilt angles of 0, 90, and 180 degrees were excluded to discourage participants from recoding the stimuli into categorical codes (e.g., “vertical” or “horizontal”).

Data Collection & Preprocessing—The collection and preprocessing of MRI data was identical to the procedures described for Experiment 1. Four blocks of the Phase 1 task were obtained, each consisting of 18 trials (6 trials from each stimulus category) lasting 6 min 56 s, for a total of 27 min 44 s in functional scans. In the same scanning session, eight blocks of the Phase 2 task were also obtained, each consisting of 9 trials lasting 7 min 14 s, for a total of 57 min 52 s in functional scans. Across both tasks, each participant performed memory tasks for a total of 85 min 19 s. A feature selection ANOVA was applied to the training data, as in Experiment 1, to remove uninformative voxels. The average number of voxels selected across participants was 11,184 (s.d. 2,648). Voxels from these masks served as input nodes to the pattern classifier for hypothesis testing.

Multivariate Pattern Analysis: Training—A pattern classifier was trained, separately for each participant, on data from the delay-period of the Phase 1 task. Data from the final 6 s of the 7.5-sec delay period, at intervals of TR = 2 s, were used to train a classifier to distinguish patterns of brain activity corresponding to the short-term retention of information encoded primarily in a phonological (pseudoword trials), semantic (word trials), or visual (line trials) form. As in Experiment 1, all data were shifted back in time by 4 s to account for hemodynamic lag of the BOLD signal. Therefore, the 6 s of fMRI data that were used from each trial were actually data that were recorded between 8 and 14 s after the beginning of the trial. To improve the interpretability of the whole-trial decoding of the Phase 2 data, we also trained the classifier on resting state brain activity during the unfilled inter-trial interval (ITI). Resting activity served as a “ground reference” for the classifier, analogous to how the Earth serves as a zero-voltage ground reference for electrical circuits. Training the classifier with rest activity did not alter the classifier’s assessment of the relative differences between the three stimulus categories during the task-portion of the trial. It did, however, normalize the classifier’s assessment such that evidence for the stimulus categories was low during the rest periods, consistent with the fact that participants were not performing a memory task during these periods of the experiment. Data from the ITI was randomly sampled so that, within each block of trials, the classifier was trained on the same number of exemplars for all four categories (72 total TRs each of phonological, semantic, visual, and ITI across the whole experiment). A unique classifier was created for each participant and applied only to that participant’s data. Classifier training accuracy was assessed and voxel importance maps (thresholded at an importance value of 0.075) were calculated as described above for Experiment 1.

Classification for Experiment 2 was carried out using penalized logistic regression, using L_2 regularization with a penalty parameter of 50. Regularization prevents over-fitting by punishing large weights during classifier training (Duda, Hart, & Stork, 2001). Note: classification for both Experiments 1 and 2 was initially carried out using backpropagation (see Methods, Experiment 1), but was also re-run using penalized logistic regression. Classification performance for Experiment 1 did not change (and therefore we report the initial results), but classifier performance was significantly improved for Experiment 2. We believe that L_2 regularization was particularly important for Experiment 2 because the classifier was also trained on resting state activity between trials, and therefore it partially learned to discriminate the three task conditions based on features that were in common to the three stimulus categories. Over-fitting was less problematic in Experiment 1, because the classifier was not trained on resting activity (the inter-trial intervals were filled with an arithmetic task).

Multivariate Pattern Analysis: Testing—A pattern classifier for each participant, trained on all four blocks of Phase 1 data, was used to assess the extent to which category-specific patterns of brain activity reappeared during the Phase 2 task. Preprocessed fMRI data at intervals of TR = 2 s were classified from every trial. Because the classifier was also trained on resting state activity, the evidence values at the beginning and end of each trial for the three stimulus categories were equally low, but non-zero. For display purposes, this low-level evidence was removed from all classifier evidence values so that the trial-averaged decoding traces would begin at zero. The continuous decoding of data from the entirety of the trials allowed for a complete characterization of the evolution of brain states corresponding to category-specific information inside and outside the focus of attention. If sustained brain activity reflected the contents of the focus of attention, but not all of STM, one would expect that the category information decoded by the classifier would track only that information which is held in the focus of attention. During the initial delay period, both memory items would be maintained in the focus because both were potentially relevant for the first response. Following the first relevance cue, removal of task-irrelevant information from the focus would be indicated by an attenuation of classifier evidence for that memory item. Whether the strength of classifier evidence were to drop to an intermediate level, or to baseline, would have implications for what it means for information to be in "in STM" but outside the focus of attention. On switch trials, retrieval of information from "activated LTM" back into the focus of attention would be indicated by the restrengthening of classifier evidence for the memory item cued as relevant for the second decision. In contrast, if sustained brain activity reflected the full contents of STM, we would expect that, regardless of cueing, evidence for the categories of both memory items should be detected by the classifier throughout the trial (at least until the second cue, because both stimuli had to be remembered up to that point).

Results (Phase 1)

Behavior—The mean accuracy and response time across all participants in the Phase 1 task were 94% (SEM 1) and 933 ms (SEM 22). Response times from trials with an incorrect response were excluded. A three-way repeated measures ANOVA on response accuracy with stimulus category (*phonological*|*semantic*|*visual*) as a within-subjects factor revealed a significant main effect of stimulus category ($F(2,16)=4.06$, $p=0.037$), and follow-up pairwise comparisons (with Bonferroni correction) indicated that the accuracy on semantic trials (98 %, SEM 1) was reliably higher ($p=0.037$) than the accuracy on phonological trials (91%, SEM 3). An identical ANOVA on response time revealed a significant main effect of stimulus category ($F(2,16)=4.11$, $p=0.036$), but follow-up pairwise comparisons (both with or without Bonferroni correction) indicated that there were no reliable differences between any pair of stimulus categories.

MVPA—Brain data from all Phase 1 trials were used to train a classifier separately for each participant. Group-averaged classification performance showed that brain activity from the retention interval of the Phase 1 task was reliably classified as matching the stimulus category of the trial (Fig. 5A). The classifier's prediction accuracy for each category was significantly above chance (25%) based on one-tailed, independent-sample t-tests across participants, with $p < 0.01$. The mean classifier evidence for each category showed strong category selectivity (e.g., the *phonological* classifier evidence was selectively high for phonological trials), supported by a significant interaction of trial type (*phonological/semantic/visual/ITI*) × evidence type (*phonological/semantic/visual/ITI*) from a 4×4 repeated measures ANOVA on the classifier evidence values ($F(9,72)=66.14$, $p<0.001$). Because each stimulus category putatively required short-term retention in one primary domain of representation, this result indicates that the classifier successfully differentiated visual from phonological (Baddeley, 1986) from semantic (Haarmann & Usher, 2001; Martin et al., 2003; Shivde & Thompson-Schill, 2004; Cameron et al., 2005) STM, and all three from the resting state activity recorded during the ITI. A distributed network of voxels throughout the brain was identified as important for supporting the classification of each category of stimulus (Fig. 5B).

Results (Phase 2)

Behavior—The mean accuracy and response time across all participants in the Phase 2 task were 91% (SEM 1) and 936 ms (SEM 10). Response times from trials with an incorrect response were excluded. A 2×2×6 repeated measures ANOVA on response accuracy with cue type (*repeat|switch*), probe type (*first/second*), and stimulus type (6 pairwise combinations of *phonological, semantic, & visual*) as within-subjects factors revealed a significant main effect of cue type ($F(1,8)=27.18$, $p<0.001$), indicating that participants were more accurate on repeat trials (93%, SEM 1) than on switch trials (88%, SEM 1). The main effect of probe type was non-significant ($F(1,8)=0.30$, $p=0.597$), but the main effect of stimulus type was significant ($F(5,40)=3.80$, $p=0.007$), and follow-up pairwise comparisons (with Bonferroni correction) indicated that participants responded less accurately to *visual-phonological* trials (i.e., trials in which the first stimulus that was cued was a *line*, and the other stimulus presented at the beginning of the trial was a *pseudoword*) (84%, SEM 2) than to both *phonological-semantic* trials (94%, SEM 2; $p=0.022$) and *semantic-phonological* trials (94%, SEM 2; $p=0.013$). Finally, the probe type × stimulus type interaction was significant ($F(5,40)=2.56$, $p=0.042$), and the three-way interaction of cue type × probe type × stimulus type was also significant ($F(5,40)=3.94$, $p=0.005$). An identical 2×2×6 repeated measures ANOVA on response times revealed a significant main effect of cue type ($F(1,8)=7.86$, $p=0.023$), indicating that participants were faster to respond on repeat trials (924 ms, SEM 14) than on switch trials (948 ms, SEM 15), and a significant main effect of probe type ($F(1,8)=25.23$, $p=0.001$), indicating that participants were faster to respond to the second probe (898 ms, SEM 13) than to the first probe (975 ms, SEM 15). All two-way interactions and the three-way interaction were non-significant.

MVPA—Brain data from every time point in all Phase 2 trials were decoded, separately for each participant, using a classifier that was trained on specific time points (i.e., delay and rest periods) from all Phase 1 trials. Group-averaged classification results for both repeat and switch trials (Fig. 6) revealed an initial rise in classifier evidence for all three categories in concert with the onset of the trial, although the waveforms quickly diverged as a function of whether or not the category was relevant on that trial. Classifier evidence for the trial-irrelevant category (say, for phonological information on trials that presented lines and a noun) quickly peaked at a low level and sustained this in a tonic manner until the end of the trial. The waveform deviated from this square wave-like shape only for slight increases corresponding to the onset of the two probes. Thus, the irrelevant category provided a

baseline reference against which we could quantitatively assess evidence for representation of trial-relevant information. In all trials, classifier evidence for both trial-relevant categories rose precipitously at trial onset, and remained at the same elevated level until the onset of the first cue. This indicated that both items were encoded and sustained in the focus of attention across the initial memory delay, while it was equiprobable that either would be relevant for the first memory response. Following onset of the first cue, however, classifier evidence for the two memory items diverged. Post-cue brain activity patterns were classified as highly consistent with the category of the cued item, while evidence for the uncued item dropped precipitously, becoming indistinguishable from the classifier's evidence for the stimulus category not presented on that trial (i.e., not different from baseline). If the second cue was a repeat cue, classifier evidence for the already-selected memory item remained elevated, and that of the uncued item remained indistinguishable from baseline (Fig. 6, Repeat). If, in contrast, the second cue was a switch cue, classifier evidence for the previously uncued item was reinstated, while evidence for the previously cued item dropped to baseline (Fig. 6, Switch).

A 2×3×10 repeated measures ANOVA on classifier evidence values from the first half of all trials (prior to the onset of the second cue) with cue type (*repeat*/*switch*), stimulus type (*cued*/*other*/*irrel*), and time (*TRs 1–10*) as within-subjects factors revealed a significant interaction of stimulus type × time ($F(18,144)=23.71$, $p<0.001$), confirming the validity of the pairwise comparisons between classifier evidence values (shown at the top of each graph in Fig. 6 for every 2-sec time interval) which indicate strong evidence for both memory items after encoding, followed by selective evidence for the cued item after the first cue. The three-way interaction of cue type × stimulus type × time was non-significant ($F(18,144)=0.37$, $p=0.991$), indicating that there was no discernible difference between classifier evidence for repeat and switch trials prior to the second cue (confirming that the task demands of both trial types were identical up to this point). To assess the impact of the second cue on classifier evidence, a 2×3×13 repeated measures ANOVA was performed on the classifier evidence values from the second half of the trials (posterior to the onset of the second cue) with cue type (*repeat*/*switch*), stimulus type (*cued*/*other*/*irrel*), and time (*TRs 11–23*) as within-subjects factors. Unlike the results from the first half of the trials, this analysis revealed a significant three-way cue type × stimulus type × time interaction ($F(24,192)=25.42$, $p<0.001$). This analysis confirms that repeat and switch cues had different effects on the classifier's assessment of brain activity following the second cue, such that the classifier identified persistent evidence only for the item that was cued for the second response.

Discussion

Together, these results suggest that, across the 8-sec delay periods, only the immediately behaviorally-relevant STM item, putatively in the focus of attention, was supported by persistent patterns of category-specific delay-period activity. Notably, classifier evidence for the uncued category did not maintain an intermediate level of activation, despite the fact that it remained "in" STM. One explanation for this finding, consistent with the results from Experiment 1, is that only information that is in the focus of attention is held in an active state. An alternative explanation is that the representation of the uncued stimulus may not have *disappeared*, but rather it *changed* following the cue. A related possibility is that item-specific representations (to which our category-specific classification methods were insensitive) may have survived despite the loss of category-level representations. We believe that these alternatives are unlikely because no current theories, to our knowledge, allow for the instantaneous, contextually-dependent recoding of STM representations into some alternate form of active representation (including a form devoid of category information). Nonetheless, we tested the first of these alternatives by running a follow-up analysis in

which we trained and tested a classifier with only post-cue brain activity from the Phase 2 task. K-fold cross-validation (k=8; see Methods) was used so that the classifier was trained and tested on separate data. The subset of fMRI data used for this analysis consisted of the 3 TRs (trial time = 10–16 s) following the onset of first cue from all trials. As in the original analysis, the data were shifted by 4 s to account for hemodynamic lag. For each trial, the brain data were labeled according to the category of the uncued stimulus (e.g., if the *word* was cued on a *semantic-visual* trial, the data would be labeled *visual*). Across all participants, this classification analysis failed to produce above-chance decoding of the category of the uncued stimulus. Although a null result, this finding indicates that in our data an alternative state of active representation of the uncued stimulus, if it was to exist, could not be readily identified using the same measurement and analysis techniques that successfully identified the active representation of cued stimulus.

An important question to consider when evaluating the results from Experiment 2 is how to interpret the baseline, which we operationalized as the classifier's estimates for trial-irrelevant information (i.e., for the stimulus category that wasn't presented in the trial). A likely explanation is that this low-level of elevated classifier evidence reflects a task set, or context, that is not specific to any trial stimulus, but is engaged with the onset of each trial, and disengaged at the offset. This idea is compatible with accounts of proactive interference (e.g., Gardiner, Craik, & Birstwistle, 1972). It may be that the classifier identified activity corresponding to the trial-irrelevant category because neural representations of stimuli from that category (which were presented in previous trials) were incidentally reactivated at the beginning of each trial. This possibility arises from the assumption that memory is accomplished in part by associating stimulus items to their encoding contexts (Howard & Kahana, 2002; Nairne, 2002; Sederberg, Howard, & Kahana, 2008; Polyn, Norman, & Kahana, 2009). Accordingly, when a context representation is activated – either to associate a new item to it or to retrieve an old item from it – this reactivation leads also to the reactivation of other items associated to it, and to some degree also to the reactivation of items associated to similar contexts. This process could provide an explanation for a key piece of evidence for the idea of "activated LTM" in the embedded-component model: Recognition probes matching the uncued contents in the modified Sternberg task (Oberauer, 2001), or matching list elements from recent previous trials (so-called "recent negative lures", Monsell, 1978; D'Esposito, Postle, Jonides, & Smith, 1999) are harder to reject than novel probes not encountered during the last few trials. The difficulty with rejecting this kind of lure might not come from persistent activation of their representations in LTM, but from their reactivation by the current retrieval context, which overlaps substantially with the context in which they have last been experienced.

GENERAL DISCUSSION

How does the brain retain information across brief periods of time? The embedded-component framework (Cowan, 1995; Ericsson & Kintsch, 1995; Oberauer, 2002) suggests a distinction between retention within the focus of attention and retention outside the focus in a presumably activated state of LTM. Although a link between attention and STM has been widely acknowledged for some time, the importance of internally-directed attention for selecting subsets of information *within* STM (Cowan, 1988; Bays & Husain, 2008; Chun, Golomb, & Turk-Browne, 2011) has only recently been recognized by neuroscience researchers (Griffin & Nobre, 2003; Woltz & Was, 2006; Nee & Jonides, 2008; Esterman, Chiu, Tamber-Rosenau, & Yantis, 2009; Nee & Jonides, 2011, Chun, 2011; Cowan, 2011; Gazzaley, 2011; Ikkai & Curtis, 2011; Lepsien, Thornton, & Nobre, 2011; Olivers & Eimer, 2011; Stokes, 2011; Vandenbroucke, Sligte, & Lamme, 2011). The present study provides converging neurophysiological evidence for the distinction of two states of representations within STM by demonstrating that the moment-to-moment information content of delay-

period activity reflects items in the focus of attention, but not those retained in memory outside the focus. Intriguingly, this was true whether the information in the focus was stimulating sensory receptors (as in Experiment 1), or, instead, was itself already in STM (as in Experiment 2).

Attention and memory were unconfounded by causing either an external shift of attention to trial-irrelevant stimuli or by causing an internal shift to a subset of information already being remembered. Experiment 1 showed that, following the presentation of trial-irrelevant stimuli during a delay period, ongoing brain activity carried information about the distractors on the screen, and therefore presumably in the focus of attention, and not about the items that were not on the screen but yet retained in memory (as verified by near-perfect recognition performance). One possibility is that our analysis methods were insufficiently sensitive to detect unattended STM representations in the presence of perceptual distraction. An alternative, however, is that sustained delay-period activity reflects only that information which is currently in the focus of attention rather than the full contents of STM. Experiment 2 provided evidence for the latter interpretation. It demonstrated that temporarily irrelevant items in STM were quickly removed from the focus of attention to a point at which their signature in ongoing brain activity vanished completely. However, these items could re-enter the focus, and have their active neural signature restored, if they were cued as relevant for behavior a few seconds later. These results, therefore, support the distinction between two functional states of representations in STM: inside and outside the focus of attention (Cowan, 1995; Oberauer, 2002). Both serve STM, but only representations inside the focus are detectable in the moment-to-moment patterns of delay-period brain activity.

We will now discuss in more detail a series of concerns, methodological and theoretical, that relate to our present findings. Assuming that ongoing neural activity is accompanied by a correlated pattern of regional cerebral blood flow, there are two classes of explanation for the finding of STM (outside the focus of attention) without persistent neural activity. The first is methodological. The short-term retention of information may have been accomplished via a reduced level of sustained firing that was not detectable with our fMRI protocol. A related possibility is that retention was supported by some other type of sustained activity to which BOLD is less sensitive, such as coherent low-frequency oscillations among task-specific neural populations. Note, however, that MVPA is much more sensitive than traditional measures of BOLD (Kriegeskorte et al., 2006; Norman et al., 2006). This is seen, for example, in the ability to recover stimulus-related information in V1 during the delay period of STM tasks despite the absence of above-baseline activity (Serences et al., 2009; Harrison & Tong, 2009), and in the ability to discriminate patterns of activity representing individual faces (Kriegeskorte et al., 2007). Also, the results from Experiment 2 in the present study demonstrate that the classification procedure was sensitive enough to detect superimposed patterns of brain activity corresponding to the active representations of two memory items from different stimulus categories. It was only after one item was cued during the delay period that the classifier's evidence for the items diverged.

Another methodological concern is that the experimental design may have been too insensitive to test our hypotheses. Training a classifier on brain activity from one task and using it to decode brain activity from another task (with a different set of task demands) may not succeed if the STM representations were qualitatively different as a result of the different cognitive demands of the two tasks. However, the successful detection of task-relevant stimulus representations in distraction-absent trials (Experiment 1) and in pre-cue delay periods (Experiment 2) validates that patterns of active stimulus representations were similar across the training and testing phases in each experiment. The possibility that the qualitative form of active representation changed, rather than disappeared, for items outside

the focus of attention seems unlikely, and is not anticipated by any existing theories with which we are familiar. Further, a follow-up analysis in Experiment 2 that considered brain activity from only the Phase 2 task failed to find evidence for an alternative form of active representation of the unattended memory items.

A second class of explanation for our results arises from an alternative to activation-based accounts of short-term retention. One mechanism that could accomplish short-term retention without persistent activity is weight-based retention via changes in synaptic potentiation. During the delay period the memory traces are not actively maintained in the sense of elevated firing rates or metabolic demands. Rather, they are passively retained by short-term increases in the strength of synaptic connections between neurons that represent the information. Synaptic weights can be temporarily modified via transient elevation of the concentration of presynaptic calcium ions (Mongillo, Barak, & Tsodyks, 2008), or by GluR1-dependent short-term potentiation (Erickson, Maramara, & Lisman, 2010). The information coded in these synaptic weight changes can be translated back into active neural firing if the memory is later reactivated by a retrieval cue (Nairne, 2002).

The idea that memory representations can be reactivated during short-delay tests of STM is anticipated in neural-network models of serial order recall (Burgess & Hitch, 1999; Farrell & Lewandowsky, 2002; Botvinick & Plaut, 2006; Burgess & Hitch, 2006), and in *retrieved context* models of memory search (Howard & Kahana, 2002; Sederberg et al., 2008; Polyn et al., 2009). These models suggest an interaction between two cognitive representations: a representation of the memory item, and a representation of the encoding context. These two representations can influence one another through synaptic weight changes in bidirectional associations between the item and its context. When an item is studied, an episodic memory is formed by linking the item features to the currently active pattern of contextual activity. The associations formed on the context-to-item weights allow the context representation to serve as a retrieval cue: If a particular context representation is reactivated, it can then be used to revive the item representation(s) that co-occurred with that particular context state. The reverse interaction, driven by the item-to-context associations, provides retrieval of the context representation that prevailed when that memory item was originally encountered. This latter process, described as mental time travel (Tulving, 2002), is crucial for the perpetuation of the free recall process, but is incidental to the cued recall process required by many tests of STM. Although these models arose in an attempt to explain variability in free recall performance, our present findings suggest that the memory retrieval mechanisms that they propose may also provide valid explanations for variability in cued recall performance at short memory delays.

Another objection that could be raised against our conclusions is that they seem to be contradicted by the findings of sustained activity observed with electrophysiological recordings from individual neurons in monkeys, the loss of which has been thought to indicate disruption of STM (Miller, Li, & Desimone, 1993; Miller & Desimone, 1994; Miller, Erickson, & Desimone, 1996). In contrast, our results from multivariate pattern analysis of fMRI recordings in humans indicate that persistent neural activation is not required for STM. One way to reconcile the two sets of findings is to appeal to the assumption that contents of STM are represented in the brain by highly distributed and overlapping patterns of activity (e.g., Haxby et al., 2001). Thus, the activity of individual neurons is unlikely to accurately reflect a representation that only exists in the distributed pattern of activity across many neurons. A second consideration is that these previous studies confounded attention and STM, such that the information to be remembered was the most task-relevant information throughout the memory interval, and therefore was likely to be continuously attended to. The persistent activity of individual neurons, which correlates

with performance in STM tasks, might instead reflect sustained attention, a reinterpretation which would be consistent with the present results.

The suggestion that LTM mechanisms support performance during a test of short-term retention is not novel. In dual-store models (Waugh & Norman, 1965; Atkinson & Shiffrin, 1968), the contribution of LTM is thought to supplement (and not replace) a STM system that is capable of holding several items. Neural evidence for this idea comes from neuroimaging and neuropsychological studies which have demonstrated that medial temporal lobe structures (known to be essential for LTM) also contribute to performance on tests of short-term retention (Olson, Page, Moore, Chatterjee, & Verfaellie, 2006; Olson, Moore, Stark, & Chatterjee, 2006; Hannula, Tranel, & Cohen, 2006; Nichols, Kao, Verfaellie, & Gabrieli, 2006; Jeneson, Mauldin, & Squire, 2010; Jeneson, Mauldin, Hopkins, & Squire, 2011). All theories of STM assume a capacity of more than one item, and typical estimates are around four (Luck & Vogel, 1997; Cowan, 2000). In the present study, we deliberately held the overall memory load so small (2 items maximum) that the capacity limits of STM would not be exceeded. Therefore, based on the ubiquitous assumption that sustained activity is the neural correlate of maintenance in STM, we would expect to observe persistent neural representations for all memory items in our tasks. However, our results demonstrate that only the item in the focus of attention retained its active representation during the delay period. In Experiment 2, the focus of attention demonstrably held two items at the same time, as shown by high classifier evidence for both memory items after encoding, so it was not for lack of attentional capacity that only one representation was actively represented after the cue. Rather it was the behavioral relevance of the memory item that determined its activity status.

The present research was motivated by the family of *embedded-component* theories of STM, which characterize the system enabling the short-term retention of information as consisting of a central component of STM (referred to here as the focus of attention) and a more peripheral component (commonly referred to as “activated LTM”). However, the results that we have presented here suggest that the label for the retention of information outside the focus of attention might be a misnomer – it is perhaps more accurately labeled “prioritized LTM” because this information is prioritized (i.e., it affects ongoing processing more strongly than does dormant information in LTM) but it is not supported by an active neural trace. The present study makes two important contributions to the further refinement of these theories: (1) It provides some of the first evidence (see also Nee, Berman, Moore, & Jonides, 2008; Nee & Jonides, 2008; Nee & Jonides, 2011) that the distinction between the two components, which has been proposed on the basis of behavioral evidence (Cowan, 1988; Oberauer, 2002), has a neural basis; (2) It maps the time course of the neural signature of the removal of task-irrelevant information from the focus of attention, showing that it corresponds to the time course of the behavioral signature of these processes (Oberauer, 2001; Oberauer, 2005). Independent of the embedded-component model, the present study demonstrates that the active neural signature of information held in STM can be disrupted by redirecting attention externally or internally, without sacrificing the short-term retention of that information. These results raise questions about the common view that persistent maintenance of neural activity is required for short-term retention, and support an alternative interpretation: Delay-period activity may reflect the focus of attention, rather than the contents of STM.

Acknowledgments

This research was funded by the National Institutes of Mental Health: R01 MH064498 (B.P.) and F31 MH085444 (J.L.-P.).

References

Atkinson RC, Shiffrin RM. Human memory: A proposed system and its control processes. The Psychology of Learning and Motivation: Advances in Research and Theory. 1968; 2:89–195.

Baddeley, AD. Working memory. London: Oxford University Press; 1986. Working memory.

Bays PM, Husain M. Dynamic shifts of limited working memory resources in human vision. Science. 2008; 321(5890):851–4. [PubMed: 18687968]

Botvinick MM, Plaut DC. Short-Term memory for serial order: A recurrent neural network model. Psychological Review. 2006; 113(2):201–233. [PubMed: 16637760]

Burgess N, Hitch GJ. Memory for serial order: A network model of the phonological loop and its timing. Psychological Review. 1999; 106(3):551.

Burgess N, Hitch GJ. A revised model of short-term memory and long-term learning of verbal sequences. Journal of Memory and Language. 2006; 55(4):627–652.

Cameron KA, Haarmann HJ, Grafman J, Ruchkin DS. Long-Term memory is the representational basis for semantic verbal short-term memory. Psychophysiology. 2005; 42(6):643–653. [PubMed: 16364060]

Chun MM. Visual working memory as visual attention sustained internally over time. Neuropsychologia. 2011; 49(6):1407–9. [PubMed: 21295047]

Chun MM, Golomb JD, Turk-Browne NB. A taxonomy of external and internal attention. Annual Review of Psychology. 2011; 62:73–101.

Courtney SM, Ungerleider LG, Keil K, Haxby JV. Transient and sustained activity in a distributed neural system for human working memory. Nature. 1997; 386:608–611. [PubMed: 9121584]

Cowan N. Evovling conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. Psychological Bulletin. 1988; 104(2):163–191. [PubMed: 3054993]

Cowan, N. Attention and memory: An integrated framework. New York: Oxford University Press; 1995.

Cowan N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity (vol 23, pg 87, 2001). Behavioral and Brain Sciences. 2000; 24(3)

Cowan N. The focus of attention as observed in visual working memory tasks: Making sense of competing claims. Neuropsychologia. 2011; 49(6):1401–6. [PubMed: 21277880]

Curtis CE, D'Esposito M. Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences. 2003; 7(9):415–423. [PubMed: 12963473]

D'Esposito M, Postle BR, Jonides J, Smith EE. The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. Proceedings of the National Academy of Sciences (USA). 1999; 96:7514–7519.

di Pellegrino G, Wise SP. Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. Journal of Neuroscience. 1993; 13:1227–1243. [PubMed: 8441009]

Druzgal TJ, D'Esposito M. Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. Journal of Cognitive Neuroscience. 2003; 15:771–784. [PubMed: 14511531]

Duda, RO.; Hart, PE.; Stork, DG. Pattern classification. 2. New York: Wiley; 2001.

Erickson MA, Maramara LA, Lisman J. A single brief burst induces glur1-dependent associative short-term potentiation: A potential mechanism for short-term memory. Journal of Cognitive Neuroscience. 2010; 22(11):2530–40. [PubMed: 19925206]

Ericsson KA, Kintsch W. Long-Term working memory. Psychological Review. 1995; 102(2):211–244. [PubMed: 7740089]

Esterman M, Chiu YC, Tamber-Rosenau BJ, Yantis S. Decoding cognitive control in human parietal cortex. Proceedings of the National Academy of Sciences. 2009; 106(42):17974.

Farrell S, Lewandowsky S. An endogenous distributed model of ordering in serial recall. Psychonomic Bulletin & Review. 2002; 9(1):59–79. [PubMed: 12026954]

Fuster JM, Alexander GE. Neuron activity related to short-term memory. Science. 1971; 173(3997):652. [PubMed: 4998337]

Gardiner JM, Craik FIM, Birstwistle J. Retrieval cues and release from proactive inhibition. Journal of Verbal Learning and Verbal Behavior. 1972; 11(6):778–783.

Gauthier I, Skudlarski P, Gore JC, Anderson AW. Expertise for cars and birds recruits brain areas involved in face recognition. Nature Neuroscience. 2000; 3(2):191–7.

Gazzaley A. Influence of early attentional modulation on working memory. Neuropsychologia. 2011; 49(6):1410–24. [PubMed: 21184764]

Griffin IC, Nobre AC. Orienting attention to locations in internal representations. Journal of Cognitive Neuroscience. 2003; 15(8):1176–94. [PubMed: 14709235]

Haarmann H, Usher M. Maintenance of semantic information in capacity-limited item short-term memory. Psychonomic Bulletin & Review. 2001; 8(3):568–578. [PubMed: 11700909]

Hannula DE, Tranel D, Cohen NJ. The long and the short of it: Relational memory impairments in amnesia, even at short lags. Journal of Neuroscience. 2006; 26(32):8352–9. [PubMed: 16899730]

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. Nature. 2009; 458(7238):632–5. [PubMed: 19225460]

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001; 293(5539):2425–2430. [PubMed: 11577229]

Haxby JV, Petit L, Ungerleider LG, Courtney SM. Distinguishing the functional roles of multiple regions in distributed neural systems for visual working memory. Neuroimage. 2000; 11:380–391. [PubMed: 10806025]

Haynes JD, Rees G. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience. 2006; 7(7):523–34.

Hebb, DO. The organization of behavior: A neuropsychological theory. New York: Wiley; 1949.

Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, Evans AC. Enhancement of magnetic resonance images using registration for signal averaging. Computer Assisted Tomography. 1998; 22(2):324–333.

Howard MW, Kahana MJ. A distributed representation of temporal context* 1. Journal of Mathematical Psychology. 2002; 46(3):269–299.

Ikkai A, Curtis CE. Common neural mechanisms supporting spatial working memory, attention and motor intention. Neuropsychologia. 2011; 49(6):1428–34. [PubMed: 21182852]

Jeneson A, Mauldin KN, Squire LR. Intact working memory for relational information after medial temporal lobe damage. Journal of Neuroscience. 2010; 30(41):13624–9. [PubMed: 20943903]

Jeneson A, Mauldin KN, Hopkins RO, Squire LR. The role of the hippocampus in retaining relational information across short delays: The importance of memory load. Learning & Memory. 2011; 18(5):301–5. [PubMed: 21502337]

Kriegeskorte N, Formisano E, Sorger B, Goebel R. Individual faces elicit distinct response patterns in human anterior temporal cortex. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(51):20600–5. [PubMed: 18077383]

Kriegeskorte N, Goebel R, Bandettini P. Information-Based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(10):3863–3868. [PubMed: 16537458]

Kubota K, Niki H. Prefrontal cortical unit activity and delayed alternation performance in monkeys. Journal of Neurophysiology. 1971; 34(3):337–347. [PubMed: 4997822]

Kuhl BA, Rissman J, Chun MM, Wagner AD. Fidelity of neural reactivation reveals competition between memories. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(14):5903–8. [PubMed: 21436044]

Lepsien J, Thornton I, Nobre AC. Modulation of working-memory maintenance by directed attention. Neuropsychologia. 2011; 49(6):1569–77. [PubMed: 21420421]

Lewis-Peacock JA, Postle BR. Temporary activation of long-term memory supports working memory. Journal of Neuroscience. 2008; 28(35):8765–8771. [PubMed: 18753378]

Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. Nature. 1997; 390(6657):279–280. [PubMed: 9384378]

Martin RC, Wu D, Freedman M, Jackson EF, Lesch M. An event-related fmri investigation of phonological versus semantic short-term memory. Journal of Neurolinguistics. 2003; 16(4–5): 341–360.

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience. 2001; 24(4):167–202.

Miller EK, Desimone R. Parallel neuronal mechanisms for short-term memory. Science. 1994; 263:520–522. [PubMed: 8290960]

Miller EK, Erickson CA, Desimone R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. Journal of Neuroscience. 1996; 16(16):5154–5167. [PubMed: 8756444]

Miller EK, Li L, Desimone R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. Journal of Neuroscience. 1993; 13(4):1460–1478. [PubMed: 8463829]

Mongillo G, Barak O, Tsodyks M. Synaptic theory of working memory. Science. 2008; 319:1543–1546. [PubMed: 18339943]

Monsell S. Recency, immediate recognition memory, and reaction-time. Cognitive Psychology. 1978; 10(4):465–501.

Nairne JS. Remembering over the short-term: The case against the standard model. Annual Review of Psychology. 2002; 53:53–81.

Nee DE, Jonides J. Neural correlates of access to short-term memory. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(37):14228–14233. [PubMed: 18757724]

Nee DE, Jonides J. Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: Evidence for a 3-state model of memory. Neuroimage. 2011; 54(2):1540–8. [PubMed: 20832478]

Nee DE, Berman MG, Moore KS, Jonides J. Neuroscientific evidence about the distinction between short- and long-term memory. Current Directions in Psychological Science. 2008; 17(2):102–106.

Newman, EL.; Norman, KA. Cerebral Cortex (New York, NY : 1991). 2010. Moderate excitation leads to weakening of perceptual representations.

Nichols EA, Kao YC, Verfaellie M, Gabrieli JD. Working memory and long-term memory for faces: Evidence from fmri and global amnesia for involvement of the medial temporal lobes. Hippocampus. 2006; 16(7):604–16. [PubMed: 16770797]

Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: Multi-Voxel pattern analysis of fmri data. Trends in Cognitive Sciences. 2006; 10(9):424–30. [PubMed: 16899397]

Oberauer K. Removing irrelevant information from working memory: A cognitive aging study with the modified sternberg task. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2001; 27(4):948–957.

Oberauer K. Access to information in working memory: Exploring the focus of attention. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2002; 28:411–421.

Oberauer K. Control of the contents of working memory--a comparison of two paradigms and two age groups. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2005; 31(4):714–28.

Olivers CN, Eimer M. On the difference between working memory and attentional set. Neuropsychologia. 2011; 49(6):1553–8. [PubMed: 21145332]

Olson IR, Moore KS, Stark M, Chatterjee A. Visual working memory is impaired when the medial temporal lobe is damaged. Journal of Cognitive Neuroscience. 2006; 18(7):1087–97. [PubMed: 16839283]

Olson IR, Page K, Moore KS, Chatterjee A, Verfaellie M. Working memory for conjunctions relies on the medial temporal lobe. Journal of Neuroscience. 2006; 26(17):4596–601. [PubMed: 16641239]

Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fmri: A tutorial overview. Neuroimage. 2009; 45(1 Suppl):S199–209. [PubMed: 19070668]

Poldrack RA. Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences. 2006; 10(2):64–69. [PubMed: 16406759]

Polyn SM, Natu VS, Cohen JD, Norman KA. Category-Specific cortical activity precedes retrieval during memory search. Science. 2005; 310:1963–1966. [PubMed: 16373577]

Polyn SM, Norman KA, Kahana MJ. A context maintenance and retrieval model of organizational processes in free recall. Psychological Review. 2009; 116(1):129–56. [PubMed: 19159151]

Postle BR. Working memory as an emergent property of the mind and brain. Neuroscience. 2006; 139:23–38. [PubMed: 16324795]

Postle BR, Druzgal TJ, D'Esposito M. Seeking the neural substrates of working memory storage. Cortex; A Journal Devoted to the Study of the Nervous System and Behavior. 2003; 39:927–946.

Postle BR, Idzikowski C, Della Salla S, Logie RH, Baddeley AD. The selective disruption of spatial working memory by eye movements. Quarterly Journal of Experimental Psychology. 2006; 59:100–120.

Quamme JR, Weiss DJ, Norman KA. Listening for recollection: A multi-voxel pattern analysis of recognition memory retrieval strategies. Frontiers in Human Neuroscience. 2010; 4

Ranganath C, Cohen MX, Dam C, D'Esposito M. Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory recall. Journal of Neuroscience. 2004; 24:3917–3925. [PubMed: 15102907]

Ranganath C, DeGutis J, D'Esposito M. Category-Specific modulation of inferior temporal activity during working memory encoding and maintenance. Cognitive Brain Research. 2004; 20:37–45. [PubMed: 15130587]

Ruchkin DS, Grafman J, Cameron K, Berndt RS. Working memory retention systems: A state of activated long-term memory. Behavioral and Brain Sciences. 2003; 26:709–777. [PubMed: 15377128]

Sederberg PB, Howard MW, Kahana MJ. A context-based theory of recency and contiguity in free recall. Psychological Review. 2008; 115(4):893–912. [PubMed: 18954208]

Serences JT, Ester EF, Vogel EK, Awh E. Stimulus-Specific delay activity in human primary visual cortex. Psychological Science. 2009; 20(2):207–14. [PubMed: 19170936]

Sheth BR, Shimojo S. Signal strength determines the nature of the relationship between perception and working memory. Journal of Cognitive Neuroscience. 2003; 15(2):173–84. [PubMed: 12676055]

Shivde GS, Thompson-Schill SL. Dissociating semantic and phonological maintenance using fmri. Cognitive, Affective, & Behavioral Neuroscience. 2004; 4:10–19.

Stokes MG. Top-Down visual activity underlying VSTM and preparatory attention. Neuropsychologia. 2011; 49(6):1425–7. [PubMed: 21315093]

Takeda M, Naya Y, Fujimichi R, Takeuchi D, Miyashita Y. Active maintenance of associative mnemonic signal in monkey inferior temporal cortex. Neuron. 2005; 48:839–848. [PubMed: 16337920]

Tulving E. Episodic memory: From mind to brain. Annual Review of Psychology. 2002; 53(1):1–25.

Vandenbroucke AR, Sligte IG, Lamme VA. Manipulations of attention dissociate fragile visual short-term memory from visual working memory. Neuropsychologia. 2011; 49(6):1559–68. [PubMed: 21236273]

Vogel EK, McCollough AW, Machizawa MG. Neural measures reveal individual differences in controlling access to working memory. Nature. 2005; 438(7067):500–3. [PubMed: 16306992]

Waugh NC, Norman DA. Primary memory. Psychological Review. 1965; 72(2):89. [PubMed: 14282677]

Wickens DD. Encoding categories of words - empirical approach to meaning. Psychological Review. 1970; 77(1):1–15.

Woltz DJ. Perceptual and conceptual priming in a semantic reprocessing task. Memory and Cognition. 1996; 24:429–440.

Woltz DJ, Was CA. Availability of related long-term memory during and after attention focus in working memory. Memory & Cognition. 2006; 34(3):668–684.

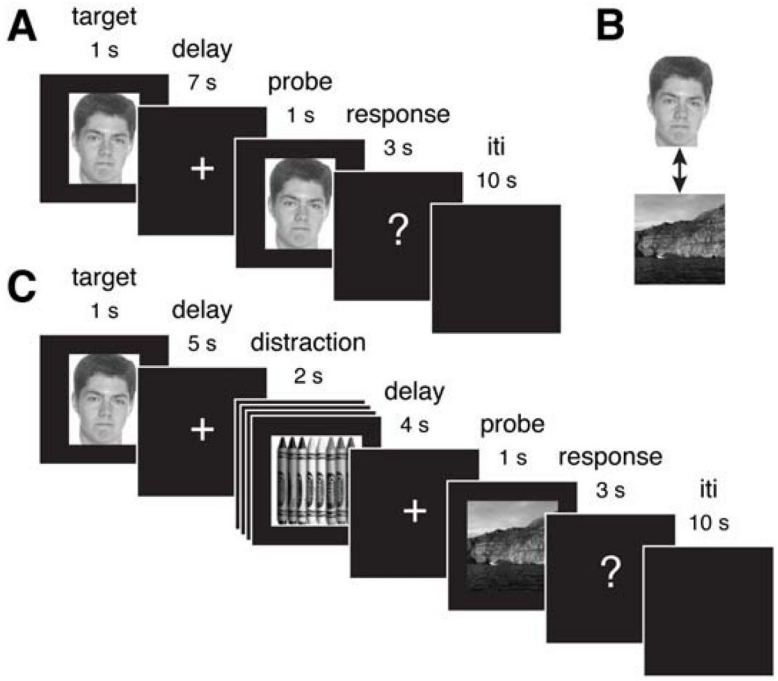

Figure 1. Task diagrams for Experiment 1

(A) In the first session, participants performed short-term recognition of faces, places, & objects inside the scanner. (B) At the beginning of the second session, outside the scanner, participants learned arbitrary cross-category pairs of stimuli. (C) Participants then returned to the scanner to perform short-term paired-associate recognition of the stimulus pairs they learned. Half of these trials included trial-irrelevant distraction during the delay period.

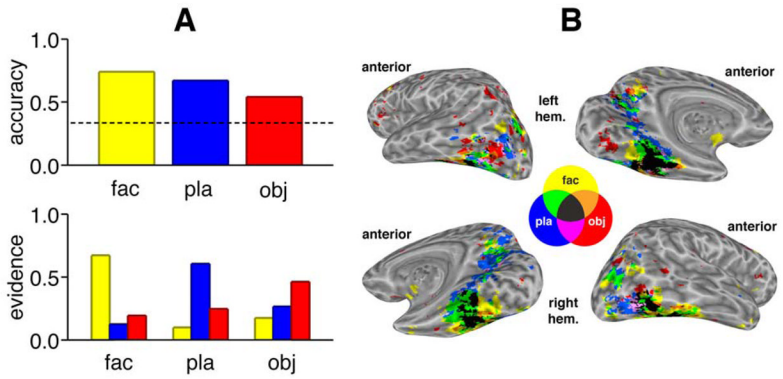

Figure 2. Classifier training for Experiment 1

(A) Classification results for the Phase 1 data is shown separately for all three categories on which it was trained: face (*fac,* yellow), place (*pla*, blue), and object (*obj*, red). Prediction accuracy is shown on the top graph (chance-level accuracy of 0.33 is indicated by the dashed line), and average classifier evidence is shown on the bottom graph. The evidence values reflect reliable category discrimination (e.g., for face trials, the classifier's evidence for *face* was much higher than its evidence for either *place* or *object*). (B) Classifier-derived voxel importance maps show voxels whose activity exerted a strong influence on the classifier's identification of a particular category. Group-averaged data are displayed on an inflated brain (*left hemisphere* in top tow, *right hemisphere* in bottom row; *lateral view* in left column; *medial view* in right column). Brain areas are colored according to the venn diagram in the center (e.g., black represents an overlap of all three categories).

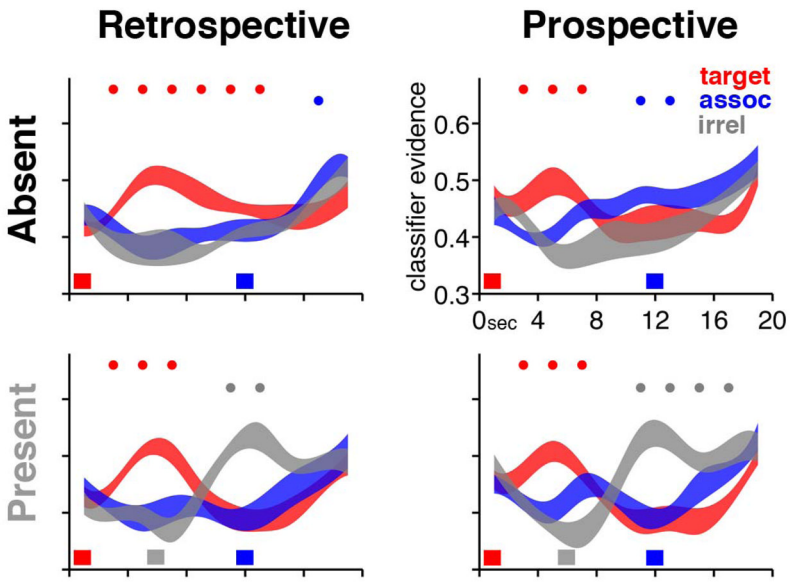

Figure 3. Classifier decoding for Experiment 1

Mean classifier evidence values are shown separately for the retrospective strategy group (left column: *Retrospective*) and the prospective strategy group (right column: *Prospective*), and separately for the distraction-absent trials (top row: *Absent*) and the distraction-present trials (bottom row: *Present*). Evidence values for the face, place, and object categories were relabeled and collapsed across all trials into three new categories: *target* (red, the category of the target stimulus on a given trial), *assoc* (blue, the category of the target's associate stimulus), and *irrel* (grey, the trial-irrelevant category). Data for each category are shown as ribbons whose thickness indicate +/− 1 SEM across participants, interpolated across the ten discrete data points in the trial-averaged data. The colored bars along the horizontal axis indicate the onset of the target (red, 0 s), the distractors (grey, 6 s; distraction-present trials only), and the probe (blue, 12 s). Statistical comparisons of evidence values for the three categories focused on within-subject differences. For every 2-sec interval throughout the trial, color-coded circles at the top of each graph indicate the category whose evidence was greater ($p<0.05$, based on repeated measures t-tests) than the average evidence for the other two categories. Unlike the data from Phase 1 that was used to train the classifier, these data were not shifted in time, and therefore the peak response to a trial event appears approximately 4 to 6 s after the onset of the event.

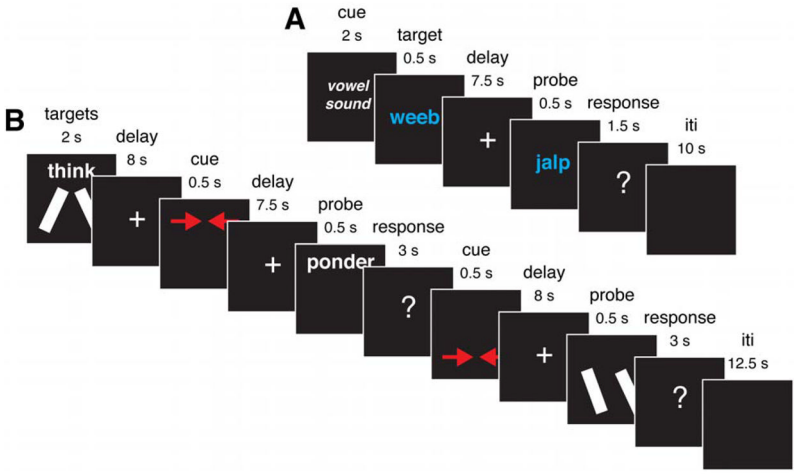

Figure 4. Task diagrams for Experiment 2

(A) In the first phase, participants performed short-term recognition of a pseudoword (phonological STM), a word (semantic STM) or two lines (visual STM). (B) In the second phase, during the same scanning session, participants performed short-term recognition with two stimuli (between-category combinations of pseudowords, words, and lines). On half of the trials, the same memory item was selected as behaviorally relevant by the first and second cues (repeat trials), and on the other half of trials the second cue selected the previously uncued item (switch trials).

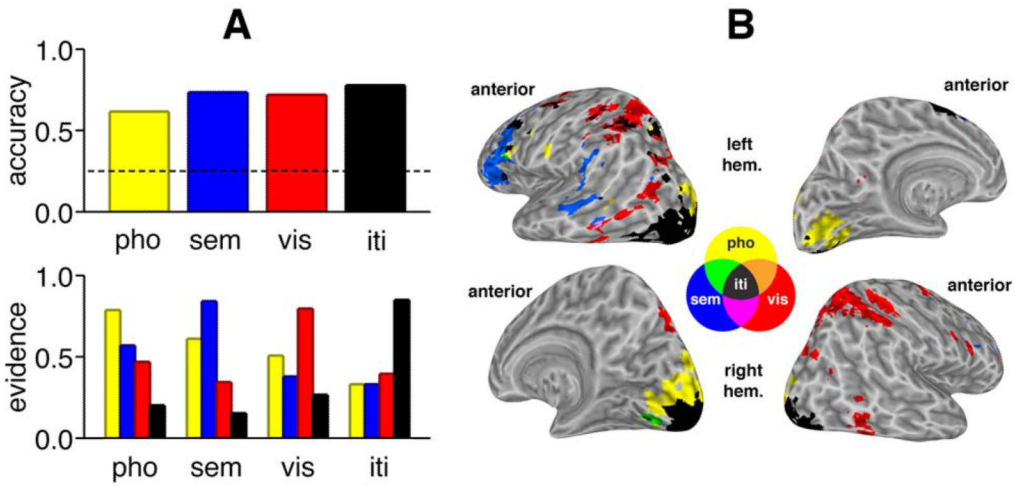

Figure 5. Classifier training for Experiment 2

Graph conventions are described in Figure 2. The classifier training performance (A) and voxel importance maps (B) are shown for phonological (*pho*), semantic (*sem*), visual (*vis*), and resting state brain activity from the inter-trial interval (*iti*). Chance-level predication accuracy was 0.25 and is indicated by the dashed line on the top graph in (A). Unlike in Figure 2, the voxels on the inflated brain hemispheres in (B) correspond to an overlap between one or more trial categories with the *iti* category. There were no important voxels that overlapped for *pho, sem,* and *vis* but not *iti,* and there were very few important voxels for the *iti* alone, and so all *iti*-related voxels were painted black.

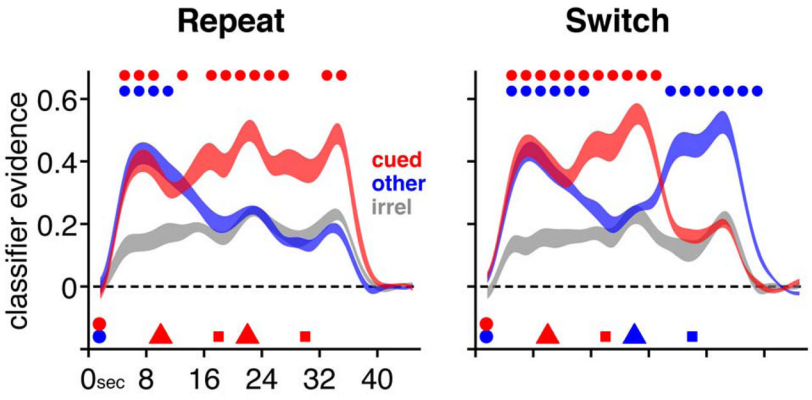

Figure 6. Classifier decoding for Experiment 2

Results are shown separately for repeat (left) and switch (right) trials. Classifier evidence values for phonological, semantic, and visual were relabeled and collapsed across all trials into three new categories: *cued* (red, the category of the memory item selected by the first cue), *other* (blue, the category of the other memory item), and *irrel* (grey, the trial-irrelevant category). The colored shapes along this horizontal axis indicate the onset of the targets (red and blue circles, 0 s), the first cue (red triangle, 10 s), the first recognition probe (red square, 18 s), the second cue (red or blue triangle, 22 s), and the final recognition probe (red or blue square, 30 s). Data for each category are shown as ribbons whose thickness indicate +/− 1 SEM across participants, interpolated across the 23 discrete data points in the trial-averaged data. Statistical comparisons of evidence values focused on within-subject differences. For every 2-sec interval throughout the trial, color-coded circles along the top of each graph indicate that the classifier's evidence for the *cued* or *other* categories, respectively, was reliably stronger ($p<0.002$, based on repeated measures t-tests) than the evidence for the trial-irrelevant category (*irrel*).